Automatic Aorta Segmentation with Transformer Based Visual Models

Jean Choi
Department of Data Science, Seoul National University
Seoul, South Korea
wisecat@snu.ac.kr

Kyu-Wha Lee
Department of Mechanical Engineering, Seoul National University
Seoul, South Korea
ksksooo@snu.ac.kr

Jae Yong Kim
Department of Computer Science and Engineering Engineering, Seoul National University
Seoul, South Korea
jykkim111@cglab.snu.ac.kr

Donguk Kim
Department of Prelimanary Medicine, Seoul National University
Seoul, South Korea
drinkuranium@snu.ac.kr

Abstract

We propose transformer-based deep learning methods for accurate and efficient aorta segmentation which is critical to the diagnosis of abdominal aortic aneurysms (AAAs). The clinical treatment of AAAs requires accurate measurement of aorta diameter. But the manual segmentation of the aorta by human radiologists is very time-consuming and shows high interobserver variability. Our deep learning model based on UNETR architecture can accurately segment aorta with and low interobserver variability. We add attention gates and inception structures to UNETR architecture which shows the state-of-the-art performance in organ segmentation. We expected that adding attention gates would help to capture important regions among a large set of voxels in a 3D image. We also expected that constructing an inception module composed of 3 convolutional pathways which capture the contextual information in axial, sagittal, and coronal planes, respectively, would enhance the model performance. However, the dice score of the UNETR model with attention gates was 81.29 and the UNETR model with inception modules was 70.24. Those scores were lower than the baseline UNETR model, 89.0. If the hyperparameters of our model had been fine-tuned with enough time, our models probably would have achieved much better performance.

1. Introduction

Abdominal aortic aneurysm (AAA) refers to an enlargement of the abdominal aorta due to a weakened aortic wall. A weakened aortic wall can involve catastrophic ruptures which result in aortic dissection. As a result, AAA causes approximately 8000 deaths per year in the UK and 15000 deaths per year in the USA [12, 29]. Ultrasonography is often the first choice for the diagnosis of AAA. But ultrasonography can have about 3 mm errors [25], so CT is widely used for precise measurement of aorta diameter and determination of future clinical treatment. For a mild AAA smaller than 5.5 cm in aorta diameter, just a rigorous follow-up is enough for patient safety. But when the aorta diameter is larger than 5.5 cm, immediate open or endovascular surgery is required [24, 22]. Therefore accurate measurement of aorta diameter through aorta segmentation is critical for the clinical decision of AAA.

In real-world clinical practice, the size of AAA is estimated by manual measurement of the maximal aortic diameter. However, it is indeed very time-consuming and often reveals high interobserver variability. This poses a seri-

ous problem in the diagnosis of abdominal aorta aneurysms. About 35% of abdominal arctic aneurysms are missed during regular check-ups [8], only late-stage AAAs that aorta ruptures impend are diagnosed [21].

In recent days, a computer-aided diagnosis system (CAD) based on machine learning techniques, especially deep learning based on neural networks, has been widely used in medical imaging. Many studies have reported that using CAD often resulted in a better accuracy compared to human experts and faster diagnostic time regardless of task: classification of positive and negative images [1], lesion detections in various organs [27], and segmentation of organs and lesions [16]. Most studies also report that CAD has a very small interobserver variability, which is an inherent feature of CAD. Consequently, aorta segmentation through CAD can solve the problems of manual segmentation.

So in this study, we applied deep learning-based CAD to aorta segmentation. Specifically, we propose a new model which is based on UNETR architecture. UNETR exploited transformer as an encoder in U-Net-like structure and achieved state-of-the-art performance in organ segmentation [14]. We introduced the attention gate and inception module to UNETR architecture for better aorta segmentation. Attention gate helps to focus on the important input data, and the inception module makes less use of computing power with better model performance. We tested those new additional features with aorta segmentation.

Our main contributions are outlined as follows:

- We present a way to increase the efficiency of CT scans in the real clinical process. In a real-world situation, a radiologist requires a long time to properly segment the region of interest which is essential for proper diagnosis. Deep learning technology has the potential to greatly enhance the efficiency of CT scans by reducing the segmentation time.

- We propose the effectiveness of the attention gate to the visual transformer model. Since a transformer can remember large contextual features of input data, the segmentation model using a transformer can be better at understanding the positional correlations between pixels. Additional attention gate boosts the understanding by highlighting important pixels among large data of pixels.

- We enhance the ability of the model by adding inception modules to the visual transformer model. 3 pathways specialized in capturing axial, sagittal, and coronal contextual information, respectively, consist of our inception module. We replace simple 3D convolutions of the original UNETR model into our inception modules for a better understanding of complex 3D images.

2. Related Work

2.1. Aorta Segmentation

One main branch of medical AI is the segmentation of various organs, lesions, and so on. Organ segmentation is an important medical task since the volume and other features that can be extracted from the segmentation can be an important marker for diagnosing diseases.

Among organ segmentation, aorta segmentation is one of the most important organ segmentation since its segmentation result can be directly used to diagnose various diseases. For example, the results of arterial wall segmentation on CT images were directly applied to diagnosis and classification of aneurysmal ascending aorta[9] and abdominal aorta [15, 18]. Aortic dissection, one of the most critical acute aortic diseases, is also diagnosed and segmented regardless of its type [5, 7, 20]. Many worsening aortic diseases accompany the calcification of the aorta. Neural networks such as Mask R-CNN have shown comparable results to human experts in measuring the calcification of aorta [13, 17]. Deep learning methods also provide a quantification of hemodynamic features such as aortic flow, peak velocity, and dimensions on MRI images [3].

2.2. Visual Transformer

The transformer was originally developed for Natural Language Processing (NLP) since NLP requires to remember large contexture features of input data [31]. Convolutional Neural Networks (CNNs) were very good at capturing spatial structure. But since CNN has a limited receptive field for each layer, its ability to capture large spatial structures is limited. Many studies have suggested that using a transformer as an encoder of the segmentation model can be a solution to this problem.

For example, ViT exploits a pure transformer applied to sequences of image patches for image classification task [10]. DeiT also utilizes Convolution-free transformer architecture with a distillation token that ensures the attention [30]. Those image classification models are widely used in segmentation models as a backbone model. Max-DeepLab uses bipartite matching and dual-path transformer and CNN for semantic segmentation [33]. VisTR conducts instance sequence matching and segmentation in video clips based on visual transformer [34].

Many models are developed for applications in the medical domain. Since many segmentation models in the medical domain use U-net or its variants, most transformer applications are also incorporated into U-net-like structures and work as an encoder. TransUNet makes hybrid use of CNN and Transformer as an encoder with cascaded upsampler [6]. Swin-Unet formulated its transformer block with multi-head self-attention and shifted window-based multi-head self-attention [4]. UCTransNet cleverly used multi-

head cross attention and channel-wise cross attention to connect encoder and decoder layer [32]. All of the models listed above have shown much better organ segmentation results than the original ViT model.

3. Method

Our model is built upon the UNETR architecure [14]. We separately added attention gates and inception modules to UNETR model and built 2 new models.

3.1. UNETR Architecture

The overview of UNETR architecture is shown in Figure 1. Like other segmentation models in the medical domain, UNETR utilizes the encoder and decoder in the U-Net-like structure. The encoder layer consists of stacked transformers which are connected to decoder layers by skip connections.

Transformer often gets its input as 1D sequence, so 3D input volume $\mathbf{x} \in \mathbb{R}^{H \times W \times D \times C}$ with resolution (H, W, D) and C input channels is flattened into 1D sequence in this model. Then the sequence are divided into uniform non-overlapping patches where $\mathbf{x}_v \in \mathbb{R}^{N \times P^3 \times C}$ where (P, P, P) is the resolution of each patches and $N = H \times W \times D/P^3$ is the length of the sequence. A linear layer project those patches into a dimensional embedding space. A 1D learnable positional embedding $\mathbf{E}_{pos} \in \mathbb{R}^{N \times K}$ is added to preserve the spatial information as follows:

$$\mathbf{z}_0 = [\mathbf{x}_v^1\mathbf{E}; \mathbf{x}_v^2\mathbf{E}; \ldots; \mathbf{x}_v^N\mathbf{E}] + \mathbf{E}_{pos} \tag{1}$$

The embedded input then pass through a stack of transformer blocks consisting of multi-head self-attention (MSA) and multilayer perceptron (MLP) sublayers as follows:

$$\mathbf{z}'_i = \mathrm{MSA}(\mathrm{Norm}(\mathbf{z}_i - 1)) + \mathbf{z}_{i-1}, \quad i = 1 \ldots L, \tag{2}$$
$$\mathbf{z}_i = \mathrm{MSA}(\mathrm{Norm}(\mathbf{z}_i - 1)) + \mathbf{z}'_{i-1}, \quad i = 1 \ldots L, \tag{3}$$

where Norm() denotes layer normalization [2], i is the intermediate block identifier, and L is the number of transformer layers. In this model, two layers with GELU activation function consist of the MLP sublayers, while the MSA sublayers have n parallel self-attention (SA) heads architecture.

SA block learns how to map between a query $\mathbf{q}$ and the corresponding key $\mathbf{k}$ which ultimately focuses on embedded sequence $\mathbf{v}$ in $\mathbf{z} \in \mathbb{R}^{N \times K}$. The attention weights A represents the similarity between $\mathbf{z}$ and their key-value pairs as follows:

$$\mathrm{A} = \mathrm{Softmax}\left(\mathbf{q}\mathbf{k}^T/\sqrt{K_h}\right) \tag{4}$$

where $K_h = K/n$ is a scaling factor for prevent the variance of the number of parameters. Each SA heads exploit the computed attention weight to compute its output as:

$$\mathrm{SA}(\mathbf{z}) = \mathbf{AV} \tag{5}$$

Then MSA sublayer compute the output as follows with its SA blocks:

$$\mathrm{MSA}(\mathbf{z}) = [\mathrm{SA}_1(\mathbf{z}); \mathrm{SA}_2(\mathbf{z}); \ldots; \mathrm{SA}_\mathrm{n}(\mathbf{z})](\mathrm{W})_{\mathrm{msa}} \tag{6}$$

where $\mathbf{W}_{\mathrm{msa}} \in \mathbb{R}^{n \cdot K_h \times K}$ denotes the multi-headed learnable weights.

Just like U-Net architecture, spatial information from multiple resolutions is stored via skip connection [26]. In the decoding layer, encoded representations are expanded through deconvolution which increases the resolution by a factor of 2. Then the information from the skip connection is concatenated. Upsized representations are then fed into $3 \times 3 \times 3$ convolutional layers twice. This deconvolution process is repeated by 4 times. The decoded data is fed into $1 \times 1 \times 1$ convolutional layer with a softmax activation functions. This merges channel-wise information and generates a voxel-wise semantic segmentation picture.

The loss function of the UNETR is a hybridization of soft dice loss and cross-entropy loss:

$$L(G, Y) = 1 - \frac{2}{J}\sum_{j=1}^{J}\left(\frac{\sum_{i=1}^{I} G_{i,j}Y_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} Y_{i,j}^2}\right) - \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} G_{i,j}\log Y_{i,j} \tag{7}$$

where I is the number of voxels; J is the number of classes; $Y_{i,j}$ is the probability of output; $G_{i,j}$ is one-hot encoded ground truth for class j at voxel i, respectively.

3.2. Attention gate

U-Net-like architectures combine intermediate feature maps from the encoder outputs via skip connection to the decoding sequence by concatenating the feature maps to decoded images and applying convolutional layers. The underlying insight is that intermediate feature maps information can indicate what is important for expanding the contracted images. This insight is similar to what attention gates aim to achieve. A limitation of convolutional layers during expanding path is that it uses only small neighboring pixels to learn how to mix the feature maps and contracted images. An attention gate can overcome the limitation of convolutional layers by learning how to use that information in the context of the entire image [23].

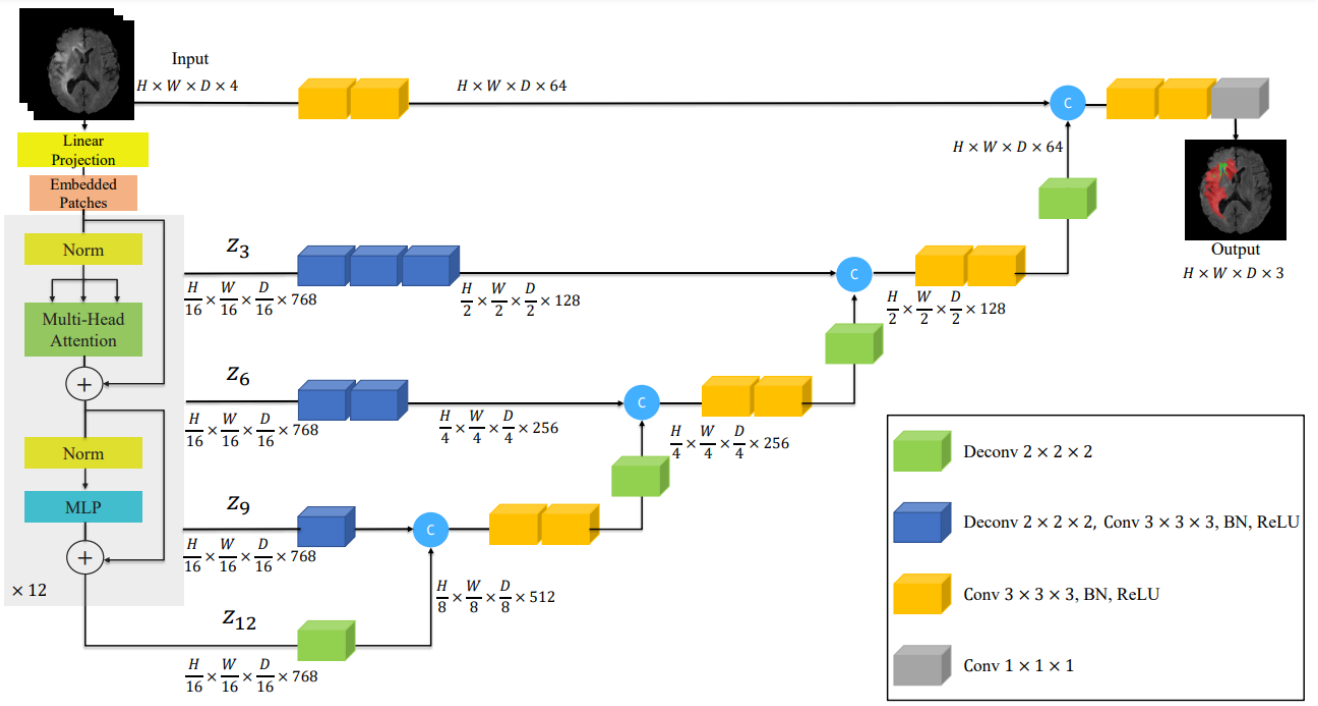

Figure 1. The overview of UNETR architecture, from [14]. The 3D input array is flattened into a 1D sequence and embedded by a linear layer with positional embedding. Then the sequence is encoded by a stack of transformers while its low-level features are stored via skip connection. The encoded sequence is then decoded through deconvolution layers with the concatenation of information stored in skip connection.

The architecture of the UNETR+attention gate model is shown in Figure 2. The attention gate aims to learn the attention coefficient $\alpha_i \in [0, 1]$. The attention coefficient indicates what pixels to focus from the input feature maps by multiplying itself to the feature maps in element-wise manner: $\hat{x}^l_{i,c} = x^l_{i,c} . ^l_i$. The contextual information is originally stored in the gating vector $g_i \in \mathbb{R}^{F_g}$. Then additive attention method is used to compute the final attention scores. Additive attention is computationally more expensive than multiplicative attention, but it is known that additive attention reveals better accuracy [19]. The formulation to compute the attention coefficient $\alpha_i \in [0, 1]$ is:

$$q^l_{att} = \psi^T \left(\sigma_1 \left(W^T_x x^l_i + W^T_g g^l_i + b_g\right)\right) + b_\psi \quad (8)$$

$$\alpha^l_i = \sigma_2 \left(q_a tt^l \left(x^l_i, g_i; \Theta_{att}\right)\right) \quad (9)$$

where $\sigma_1(x_{i,c}) = max(0, x_{i,c})$ is the ReLU activation function and $\sigma_2(x_{i,c} = \frac{1}{1+exp(-x_{i,c})}$ is the sigmoid activation function. Softmax activation function can also be considered, but sigmoid activation function is preferred since softmax activation function yields sparser activations. The parameters of the attention gates are: linear transformations $W_x \in \mathbb{R}^{F_l \times F_{int}}$, $W_g \in \mathbb{R}^{F_g \times F_{int}}$, $\psi \in \mathbb{R}^{F_{int} \times 1}$ and bias terms $b_\psi in \mathbb{R}$, $b_g \in \mathbb{R}^{F_{int}}$. The linear transformations are comptue channel-wisely through $1 \times 1 \times 1$ convolution layers.

3.3. Inception Module

Better model performance and generalizability can be obtained by ensembling multiple individual models [11]. The inception module cleverly exploited the insight of ensembling. Instead of using a single convolutional pathway, the inception module has multiple convolutional pathways. Each pathway has different convolutional layers in size, dimension, and so on, so each of them captures different features from the input images. By combining information from each pathway with a different specialty, the inception model is better to capture complex contextual information [28].

The original UNETR model applies 3D convolutional layers to capture the 3D contextual information. But simply using 3D convolutions has several problems. First, 3D convolutions are computationally expensive. A $3 \times 3 \times 3$ convolution has 3 times more parameters than a 3×3 convolution. Since dot product is applied to entire 3D images and 3D convolutions, the number of operations in 3D convolutions soars compared to 2D convolution. Second, 3D convolutions may not be enough to capture the complexity of 3D images which have much more entanglements between

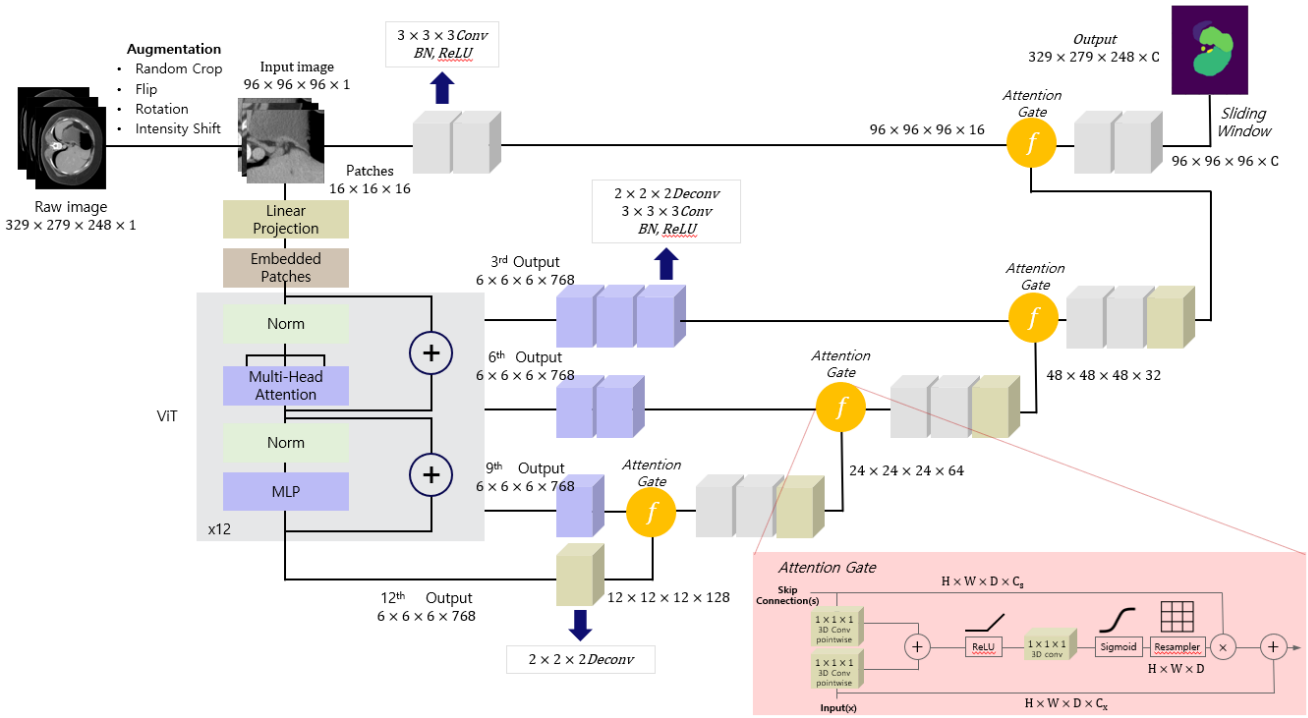

Figure 2. The overview of UNETR+attention gate architecture. The original UNETR model uses deconvolution with concatenated feature map outputs from encoder transformer layers. By using attention gates instead of just concatenating feature maps, the model can capture more important pixels which are helpful to identify a clearer context.

voxels. But the computational limitation of 3D convolutions puts a limit on model depth.

We tried to overcome those problems by using an inception module instead of a single 3D convolution. The overall architecture of our model is shown in Figure 3. The structure of our inception module is outlined in Figure 4. To reduce the computational overhead, input from the previous layer is processed by 1 convolutions. Then 3 paths examine axial, sagittal, coronal planes by applying 2D convolutions, respectively. Each path learns 3D image features by applying an additional 1D convolutions. All of those paths are concatenated at the last step. We replaced 3D convolutions in UNETR model with inception modules, as shown in Figure 3. We also applied inception modules in skip connections for the better composition of feature maps.

4. Experiment

4.1. Dataset

We applied our model to a data set from 'Beyond the Cranial Vault' (BTCV) segmentation challenge. The abdominal CT images of the BTCV data set were acquired at the Vanderbilt University Medical Center (VUMC) for ongoing colorectal cancers chemotherapy trial or a retrospective ventral hernia study. The BTCV data set include 50 abdomen CTs with 13 organs including the aorta are manually labeled by trained radiologists or radiological oncologist. For our purpose, we only used aorta masks from the data set. Images have different slice thicknesses from 2.5 mm to 5.0 mm. The distribution of in-plane resolution is also various from $0.54 \times 0.54\ mm^2$ to $0.98 \times 0.98\ mm^2$ with variable volume sizes ($512 \times 512 \times 85$ – $512 \times 512 \times 198$) and field of views (approx.$280 \times 280 \times 280\ mm^3$ – $500 \times 500 \times 650\ mm^3$).

4.2. Evaluation Metric

We compared the accuracy of segmentation to the reference standard segmentation using Dice score. For a given semantic class, let G_i and P_i denote the ground truth and prediction values for voxel i and G and P denote ground truth and prediction segmentation maps respectively. The Dice score metrics are defined as

$$Dice(G, P) = \frac{2\sum_{i=1}^{I} G_i P_i}{\sum_{i=1}^{I} G_i + \sum_{i=1}^{I} P_i} \tag{10}$$

4.3. Implementation Details

UNETR + Attention Gate model and UNETR + Inception module were trained in different environment. UN-

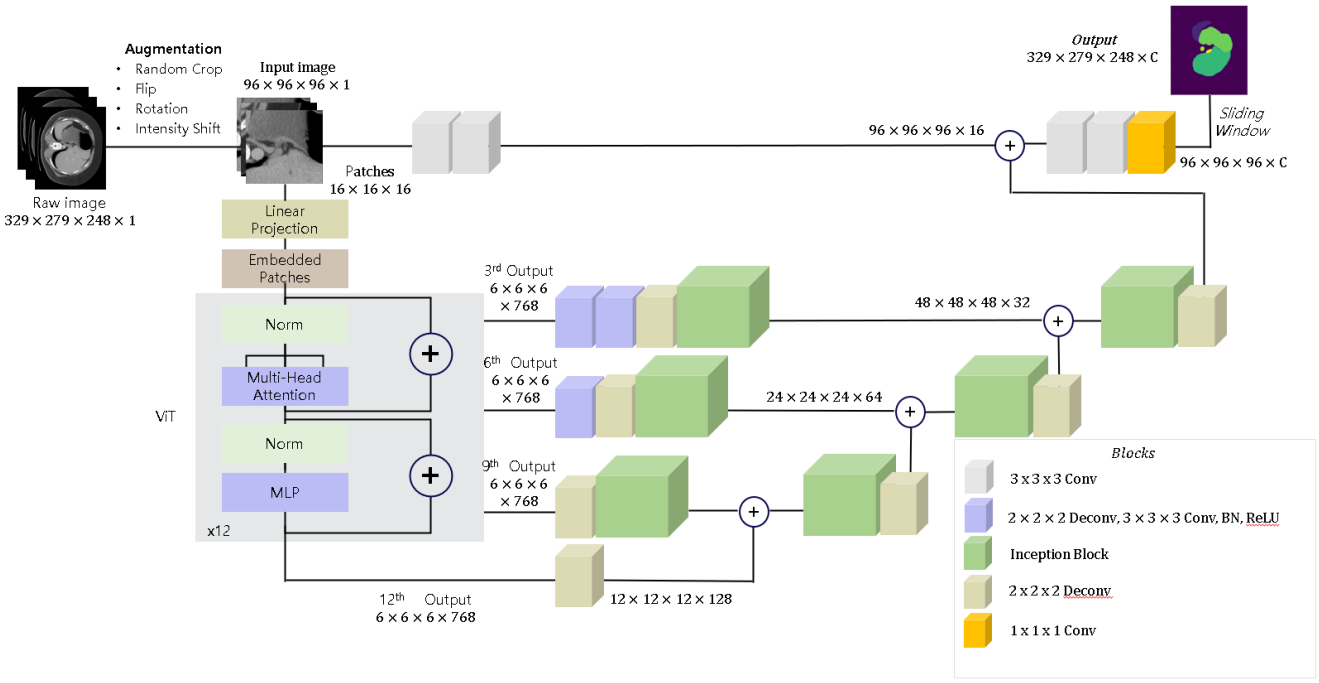

Figure 3. The overview of UNETR+inception module architecture. The original UNETR model only uses 3D convolutions to extract contextual information. As GoogLeNet proposed, models with parallel convolutional paths are better at capturing image features [28] Our inception module parallel learns image features from axial, sagittal, and coronal planes, respectively, instead of just learning pure 3D features by 3D convolutions as done in the original UNETR model.

ETR + Attention Gate : We implement this model in PyTorch and the open-source MONAI framework. The model was trained using 2 NVIDIA GTX 2080Ti graphic cards. For the attention gated model, we used UNETR pre-trained weights for transformer, deconvolution and convolution layers. Apart from the convolutional layers, the attention gates are initialized with the He normal initializaion method. Without any freezing, all parameters are fine tuned and transfer learning alleviates the computational cost.

UNETR + Inception module : we implement our model in PyTorch and the open-source MONAI framework. The model was trained using 1 NVIDIA Titan V graphic card. For this model, transfer learning was not used.

All models were trained with a batch size of 1, using the AdamW optimizer with an initial learning rate of 0.0001 and weight decay of 0.00001 for 5,000 iterations. In order to focus on the segmentation of aorta, we discarded all labels except the aorta label from the BTCV dataset. As the training dataset only contains 30 images, data augmentation was used to avoid early overfitting. For data augmentation, we used strategies such as random flip in axial, sagittal, and coronal views, random rotation of 90, 180, 270 degrees, and intensity shifts. Our transformer-based encoder follows the ViT-B16[10] architecture with L = 12 layers, an embedding size of K = 768. We used a patch resolution of 16 x 16 x 16. For training, the training dataset was split into 24 images and 6 images for the training set and validation set accordingly. The test dataset contains 20 images without any labels. For inference, we used a sliding window approach with the same resolution as our input resolution and an overlap portion of 0.8 between the neighboring patches, so that the input image will be padded when the ROI size is larger than the inputs' spatial size during inference.

Methods	Aorta Segmentation(CT)
TransUNet	0.889
CoTr	0.920
SETR NUP	0.867
ASPP	0.918
UNETR	0.890
UNETR+Inception	0.702
UNETR+AG	0.813

Table 1. Quantitative comparions of segmentation performance of aorta segmentation in the BTCV test set. All results obtained from BTCV leaderboard.

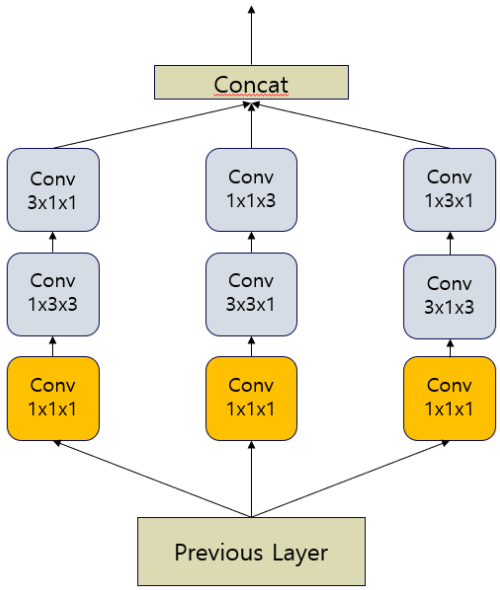

Figure 4. The detailed structure of the inception module in our model. 3 paths are specialized to capture the contexture information in axial, sagittal, and coronal planes, respectively. The module is better at resolving the complex contexture information of a 3D image than just a single 3D convolutional layer. The most left path is a 1x1x1 convolution for channel reduction.

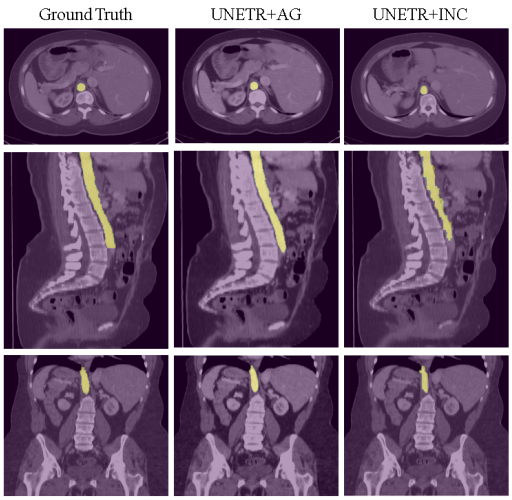

Figure 5. Qualitative visualization of outputs from our models. Axial, sagittal, and coronal views are shown for each model and ground truth.

4.4. Results and Evaluations

Our attention gate model achieves an overall average Dice score of 81.29 and our inception model achieves an average dice score of a 70.24 as shown in Table 1. Our models does not outperform the original UNETR model, but our models outperform previous CNN encoder-decoder based models such as UNet-3D and attention-UNet. This validates the effectiveness of transformers as encoders. We could not fine-tune the hyperparamteres of our model due to the time-limit, while the baseline models had been highly optimized to the competition. We believe the performance of our models will be improved after dedicated optimization of the hyperparameters. Qualitative multi-organ segmentation comparisons are presented in Figure 5. UNETR+Attention gate model shows better results overall in terms of localization of pixels around contour regions. Results from the UNETR+Inception model show that the context information learning is somewhat durable but the localization ability is not robust enough near the contour regions.

5. Conclusion

This paper introduces a transformer-based architectures for semantic segmentation of volumetric medical images by redefining this task a 1D sequence-to-sequence prediction problem. Not only to capture global contextual representation and learn long-range dependencies, but also to increase the model's capacity to focus on important parts in intermediate feature maps, we proposed attention gates and inception blocks. Although our proposed models did not outperform UNETR for aorta segmentation, we could observe the effectiveness of transformer based encoders on general volumetric medical images although the hyperparameters of the our model was not fine-turned. More discussion is needed on the usage of inception blocks in medical image segmentation as the localization around contour regions turn out to be poorer than previous models.

References

[1] Enes Ayan and Halil Murat Ünver. Diagnosis of pneumonia from chest x-ray images using deep learning. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pages 1–5. Ieee, 2019. 2

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[3] Haben Berhane, Michael Scott, Mohammed Elbaz, Kelly Jarvis, Patrick McCarthy, James Carr, Chris Malaisrie, Ryan Avery, Alex J Barker, Joshua D Robinson, et al. Fully automated 3d aortic segmentation of 4d flow mri for hemodynamic analysis using deep learning. *Magnetic resonance in medicine*, 84(4):2204–2218, 2020. 2

[4] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021. 2

[5] Long Cao, Ruiqiong Shi, Yangyang Ge, Lei Xing, Panli Zuo, Yan Jia, Jie Liu, Yuan He, Xinhao Wang, Shaoliang Luan,

et al. Fully automatic segmentation of type b aortic dissection from cta images enabled by deep learning. *European journal of radiology*, 121:108713, 2019. 2

[6] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2

[7] Junlong Cheng, Shengwei Tian, Long Yu, Xiang Ma, and Yan Xing. A deep learning algorithm using contrast-enhanced computed tomography (ct) images for segmentation and rapid automatic detection of aortic dissection. *Biomedical Signal Processing and Control*, 62:102145, 2020. 2

[8] Rachel Claridge, Sam Arnold, Neil Morrison, and André M van Rij. Measuring abdominal aortic diameters in routine abdominal computed tomography scans and implications for abdominal aortic aneurysm screening. *Journal of vascular surgery*, 65(6):1637–1642, 2017. 2

[9] Albert Comelli, Navdeep Dahiya, Alessandro Stefano, Viviana Benfante, Giovanni Gentile, Valentina Agnese, Giuseppe M Raffa, Michele Pilato, Anthony Yezzi, Giovanni Petrucci, et al. Deep learning approach for the segmentation of aneurysmal ascending aorta. *Biomedical Engineering Letters*, 11(1):15–24, 2021. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 6

[11] MA Ganaie, Minghui Hu, et al. Ensemble deep learning: A review. *arXiv preprint arXiv:2104.02395*, 2021. 4

[12] Richard F Gillum. Epidemiology of aortic aneurysm in the united states. *Journal of clinical epidemiology*, 48(11):1289–1298, 1995. 1

[13] Peter M Graffy, Jiamin Liu, Stacy O'Connor, Ronald M Summers, and Perry J Pickhardt. Automated segmentation and quantification of aortic calcification at abdominal ct: application of a deep learning-based algorithm to a longitudinal screening cohort. *Abdominal Radiology*, 44(8):2921–2928, 2019. 2

[14] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. *arXiv preprint arXiv:2103.10504*, 2021. 2, 3, 4

[15] Ho Aik Hong and UU Sheikh. Automatic detection, segmentation and classification of abdominal aortic aneurysm using deep learning. In *2016 IEEE 12th International Colloquium on Signal Processing & Its Applications (CSPA)*, pages 242–246. IEEE, 2016. 2

[16] Carson Lam, Caroline Yu, Laura Huang, and Daniel Rubin. Retinal lesion detection with deep learning using image patches. *Investigative ophthalmology & visual science*, 59(1):590–596, 2018. 2

[17] Chun Yu Liu, Chun Xiang Tang, Xiao Lei Zhang, Sui Chen, Yuan Xie, Xin Yuan Zhang, Hong Yan Qiao, Chang Sheng Zhou, Peng Peng Xu, Meng Jie Lu, et al. Deep learning powered coronary ct angiography for detecting obstructive coronary artery disease: The effect of reader experience, calcification and image quality. *European Journal of Radiology*, 142:109835, 2021. 2

[18] Jen-Tang Lu, Rupert Brooks, Stefan Hahn, Jin Chen, Varun Buch, Gopal Kotecha, Katherine P Andriole, Brian Ghoshhajra, Joel Pinto, Paul Vozila, et al. Deepaaa: clinically applicable and generalizable detection of abdominal aortic aneurysm using deep learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 723–731. Springer, 2019. 2

[19] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015. 4

[20] Tianling Lyu, Guanyu Yang, Xingran Zhao, Huazhong Shu, Limin Luo, Duanduan Chen, Jiang Xiong, Jian Yang, Shuo Li, Jean-Louis Coatrieux, et al. Dissected aorta segmentation using convolutional neural networks. *Computer Methods and Programs in Biomedicine*, 211:106417, 2021. 2

[21] Matthew W Mell, Mark A Hlatky, Jacqueline B Shreibati, Ronald L Dalman, and Laurence C Baker. Late diagnosis of abdominal aortic aneurysms substantiates underutilization of abdominal aortic aneurysm screening for medicare beneficiaries. *Journal of vascular surgery*, 57(6):1519–1523, 2013. 2

[22] Fatima S Merali and Sonia S Anand. Immediate repair compared with surveillance of small abdominal aortic aneurysms. *Vascular medicine (London, England)*, 7(3):249–250, 2002. 1

[23] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 3

[24] The UK Small Aneurysm Trial Participants. Mortality results for randomised controlled trial of early elective surgery or ultrasonographic surveillance for small abdominal aortic aneurysms. *The Lancet*, 352(9141):1649–1655, 1998. 1

[25] Denis S Quill, Mary Paula Colgan, and David S Sumner. Ultrasonic screening for the detection of abdominal aortic aneurysms. *Surgical Clinics of North America*, 69(4):713–720, 1989. 1

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[27] Sanjay Saxena, Neeraj Sharma, Shiru Sharma, SK Singh, and Ashish Verma. An automated system for atlas based multiple organ segmentation of abdominal ct images. *Journal of Advances in Mathematics and Computer Science*, pages 1–14, 2016. 2

[28] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 1–9, 2015. 4, 6

[29] MM Thompson. Controlling the expansion of abdominal aortic aneurysms. *Journal of British Surgery*, 90(8):897–898, 2003. 1

[30] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 2

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[32] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. *arXiv preprint arXiv:2109.04335*, 2021. 3

[33] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5463–5474, 2021. 2

[34] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021. 2